CLAIMS

What is claimed is:

1. A method for identifying a repeat sequence, the method comprising the steps of:

selecting a query sequence;

testing said query sequence with a redundant file;

identifying sequences in the redundant file that contain a similar sequence to a portion of the query sequence, wherein said identified sequences and said similar portion of the query sequence make up a pairwise sequence alignment;

aligning all the identified pairwise sequence alignments;

designating the right and left endpoints of each identified sequence and any intervening sequences;

identifying a position within the query sequence corresponding to each endpoint;

defining regions within the query sequence, wherein a region is a sequence between two consecutive positions matching two endpoints; and

identifying each regions having at least five sequence matches in the identified pairwise alignments as a repeat sequence.

2. A method for constructing a repeat database comprising:

selecting a query sequence;

selecting known repeat sequences;

adding known repeat sequences into a repeat sequence database;

masking said query sequence with repeat sequences in the repeat sequence database;

testing said masked query sequence with a redundant file;

identifying sequences in the redundant file that contain a similar sequence to a portion of the query sequence, wherein said identified sequences and said similar portion of the query sequence make up a pairwise sequence alignment;

aligning all the identified pairwise sequence alignments;

designating the right and left endpoints of each identified sequence and any intervening sequences;

identifying a position within the query sequence corresponding to each endpoint;

defining regions within the query sequence, wherein a region is a sequence between two consecutive positions matching two endpoints;

identifying any two successive regions having a large variance in the number of sequence matches; and

adding the sequence within the region of the two successive regions having the highest number of sequence matches into the repeat sequence database.

3. The method of claim 2, wherein the large variance in the number of sequence matches is equal to 5 or more.

4. A database product of the process of claim 2.

5. The method of claim 1 or 2, wherein said sequence is a deoxyribonucleotide sequence.

6. The method of claim 1 or 2, wherein said sequence is a ribonucleotide sequence.

7. The method of claim 1 or 2, wherein said sequences are derived from animal DNA or RNA.

8. The method of claim 7, wherein said animal is a human.

9. The method of claim 8, wherein said animal is a mouse.

10. The method of claim 1 or 2, wherein said sequences are derived from plant DNA or RNA.

11. The method of claim 10, wherein said plant is a single-cell plant.

12. The method of claim 1 or 2, wherein said sequences are derived from fungal DNA or RNA.

13. The method of claim 1 or 2, wherein said sequences are derived from DNA or RNA of a microorganism or virus.

14. The method of claim 1 or 2, wherein said sequences are derived from DNA or RNA of a single-cell eukaryote.

15. The method of claim 1 or 2, wherein said sequences are derived from synthetic man-made DNA or RNA.

16. The method of claim 1 or 2, wherein said sequences are postulated based upon amino acid sequences.

17. The method of claim 2, wherein said database is encoded in a biological medium.

18. The method of claim 2, wherein said database is encoded in a written medium.

19. The method of claim 2, wherein said database is encoded in an electronic medium.

20. The method of claim 19, wherein said electronic medium is a computer-readable medium.

21. The method of claim 20, wherein said computer-readable medium is addressable through an internet connection.

22. The method of claim 1 or 2, wherein said redundant file is a Public Domain Database.

23. The method of claim 22, wherein said Public Domain Database is GenBank.

24. The method of claim 22, wherein said Public Domain Database is dbEST.

25. The method of claim 22, wherein said Public Domain Database is TIGR.

26. The method of claim 22, wherein said Public Domain Database is SwissProt.

27. The method of claim 1 or 2, wherein sequence comparisons are carried out using a Database Search Algorithm.

28. The method of claim 27, wherein said Database Search Algorithm is BLAST.

29. The method of claim 27, wherein said Database Search Algorithm is FASTA.

30. The method of claim 27, wherein said Database Search Algorithm is Smith-Waterman.

31. The method of claim 1 or 2, wherein said sequence comparisons are carried out utilizing a Scoring Matrix Program.

32. The method of claim 31, wherein said Scoring Matrix Program is PAM.

33. The method of claim 31, wherein said Scoring Matrix Program is BLOSUM.

34. The process of Figure 2.

35. A repeat sequence product of the process of claim 1.

36. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database produced by the method of claim 2.

37. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database produced by the method of claim 2; and,

instructions for the use of said database.

38. A kit for analyzing nucleotide sequences comprising:

an electronic medium readable by a computer, said medium encoding a database produced by the method of claim 2;

instructions for the use of said database; and,

a computer.

39. An improved database of nucleotide sequences, the improvement consisting of repeat sequences containing a similar sequence to a portion of a query sequence, wherein said identified sequences and said similar portion of the query sequence make up a pairwise sequence alignment, and wherein all identified pairwise sequence alignments have right and left endpoints of each identified sequence and any intervening sequences.